

Efficient Simulation of Financial Stress Testing Scenarios with Suppes-Bayes Causal Networks

Gelin Gao
New York University
New York, USA

Bud Mishra
New York University
New York, USA

Daniele Ramazzotti
Stanford University
Stanford, California, USA

ABSTRACT

The most recent financial upheavals have cast doubt on the adequacy of some of the conventional quantitative risk management strategies, such as VaR (Value at Risk), in many common situations. Consequently, there has been an increasing need for verisimilar financial stress testings, namely simulating and analyzing financial portfolios in extreme, albeit rare scenarios. Unlike conventional risk management which exploits statistical correlations among financial instruments, here we focus our analysis on the notion of probabilistic causation, which is embodied by Suppes-Bayes Causal Networks (SBCNs), SBCNs are probabilistic graphical models that have many attractive features in terms of more accurate causal analysis for generating financial stress scenarios.

In this paper, we present a novel approach for conducting stress testing of financial portfolios based on SBCNs in combination with classical machine learning classification tools. The resulting method is shown to be capable of correctly discovering the causal relationships among financial factors that affect the portfolios and thus, simulating stress testing scenarios with a higher accuracy and lower computational complexity than conventional Monte Carlo Simulations.

Keywords

Stress Testing, Graphical Models, Causality, Suppes-Bayes Causal Networks, Classification, Decision Trees

1. INTRODUCTION

Risk management has increasingly become a central part of world finance in the past century. Quantitative risk management generally targets the risk of insolvency: namely, the depletion of capital of a trading agency to the point that the trading agency has to stop its operations. For any trading agency, its account consists of cash, stocks, bonds or other financial instruments, and net equity, where $Equity = Cash + Financial\ Instruments$. The task of quantitative risk management is to calculate the amount of equity that has to be reserved so that the net equity will not drop

to negative [8]. Depending on different financial agencies, hedge funds, banks or clearing houses, and on different financial instruments, stocks, bonds, or derivatives, the method of risk management may vary, but the central idea behind conventional risk management remains: we assess the statistical distribution of the asset (or portfolio), and estimate the worst-case scenarios, generally by methods such as Monte Carlo Simulation [17].

However, such conventional approach got discredited by the recent events leading to major financial catastrophes. For example, in the recent 2008 financial crisis, the reserves calculated the risk by using methods such as VaR (Value at Risk) [11] which proved to be painfully inadequate. Because of that, a different method had to be introduced, namely the one of *stress testing*. Stress testing refers to the analysis or simulation of the response of financial instruments or institutions, given intensely stressed scenarios that may lead to a financial crisis [5]. For example, narrowly speaking, stress testing may model the response of a portfolio when Dow Jones suddenly drops by 5%. The difference between stress testing and conventional risk management is that stress testing deliberately introduces an adversarial, albeit plausible event, which may be highly improbable but not implausible – e.g., a black swan event triggering an unforeseen scenario. Thus, stress testing must be capable of observing the response of financial instruments or institutions under extremely rare scenarios. Such scenarios may be deemed to be unlikely to be observed in conventional risk management, where the simpler system may fail to estimate a 99th percentile of the loss distribution, and subsequently leading to a claim that, with 99% confidence level, a specific portfolio will perform well, giving a false sense of security.

Recently, many different approaches have been developed to implement some form of stress testing. In terms of stress scenario generation, the most direct method is the historical one, in which observed events from the past are used to test contemporary portfolios [12]. The historical approach is objective since it is based on actual events, but it is not necessarily relevant under the present conditions, which necessitate some hypothetical methods. As an alternative, the event-based method has been proposed in order to quantify a specific hypothetical stress scenario subjectively, by domain experts, and then estimate the possible consequence of such event using macroeconomic and financial models [12]. Event-based methods rely intensively on expert judgement on whether a hypothetical event will be severely-damaging, albeit still plausible to occur. Sometime such judgement becomes difficult when the relationship between the underlying

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

xxxxxxx xxxxxxxx, xxxxxxxx

© 2017 ACM. ISBN xxx-xxxx-xx-xxx/xx...\$xx.xx

DOI: xx.xxx/xxx-x

risk factors and the portfolio is unknown. To ensure a scenario is damaging to the portfolio, a portfolio-based method has been also studied in order to link scenarios directly with the portfolio [12]. To this extent, portfolio-based methods rely on Monte Carlo Simulation to identify the movements of risk factors that stress the given portfolio most severely, however brute force Monte Carlo Simulation is computationally inefficient especially when dealing with many risk factors. To solve this problem, Rebonato et al. proposed a sampling approach based on Bayesian networks in [18].

Roadmap: This paper is organized as follow. Next section describes the theoretical foundations of our approach and, in particular, it shows how combining the expressivity of Suppes-Bayes Causal Networks together with classical classification approaches can effectively capture the dynamics of financial stress testing. Section 3 provides an algorithm for the efficient inference of SBCNs from financial data, discusses its performance in-depth and shows on realistic simulated data how our approach is preferable in comparison to the standard Bayesian methods. Section 4 concludes the paper.

2. METHOD

In this work we use Bayesian Graphical Models [9], popularly known as Bayesian networks, as a framework to assess stress testing, as previously done in this context by [18]. Bayesian networks have long been used in biological modeling such as -omics data analysis, cancer progression or genetics [2, 10, 14], but their application to financial data analysis has been rare. Roughly speaking, Bayesian networks attempt to exploit the conditional independence among random variables, whether the variables represent genes or financial instruments. In this paper we adopt a variation of the traditional Bayesian networks as done in [15, 16], where the authors show how constraining the search space of valid solutions by means of a causal theory grounded in Suppes' notion of probabilistic causation [23] can be exploited in order to devise better learning algorithms. Also, by accounting for Suppes' notion of probabilistic causation, we ensure not only conditional independence but also *prima facie* causal relations among variables, leading us to a better definition of the actual factors leading to risk. Moreover, through a maximum likelihood optimization scheme which makes use of a regularization score, we also attempt to only retain edges in the Bayesian network (graphically depicted as a directed acyclic graph, DAG) that correspond to only genuine causation, while eliminating all the spurious causes [15, 16].

Yet, given the inferred network, we can sample from it to generate plausible scenarios, though not necessarily adversarial or rare. In the case of stress testing, it is crucial to also account for rare configurations, for this reason, we adopt auxiliary tools from machine learning to discover random configurations that are both unexpected and undesired.

In this Section we expand the concept sketched above and specifically we provide a background of our framework, by describing the adopted Bayesian models and causal theories and we show how classification given an inferred SBCN can effectively guide stress testing simulations.

2.1 Bayesian networks

Informally, Bayesian networks are defined as a directed acyclic graph (DAG) $G = (V, E)$, in which each node V

represents a random variable to which is associated a conditional probability table, and each arc E models dependency relationships. The nodes induce an overall joint distribution that can be written as a product of the conditional distributions associated with each variable [9]. In this paper, without any loss of generality, we restrict our attention to Bernoulli random variables with support in $(0, 1)$. Specifically, we will consider as inputs for our analyses a dataset D of m observations over n Bernoulli variables; we refer to the next Sections for a detailed description of the meaning of such variables. More details about Bayesian networks may be found in [9].

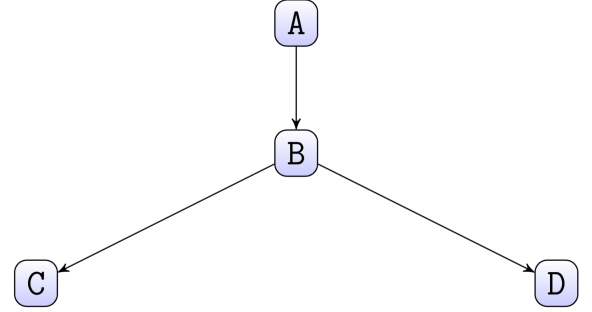


Figure 1: Example of graphical structure of a Bayesian network with 4 random variables.

Let us now consider as an example the Bayesian network shown in Figure 1, where A , B , C , and D are 4 random variables represented by four nodes, and the dependencies among the nodes are modelled by directed arcs. Loosely speaking, the link $A \rightarrow B$ indicates that the knowledge of A (the parent) influences the probability of B (the child), or A and B are statistically dependent. Furthermore, for node B , node A is called B 's parent and nodes C and D are called B 's children. More precisely, in the conditional probability tables related to the afore mentioned Bayesian network, the rows for node B specifies how the knowledge of A affects the probability of B being observed. For example, let A and B be both binary random variables with support over $(0, 1)$. Table 1 specifies the distribution of B under the condition of A , and we can see clearly the effect of the parent on the child in this example.

	$A = 0$	$A = 1$
$B = 0$	$P(B = 0 A = 0) = 0.3$	$P(B = 0 A = 1) = 0.4$
$B = 1$	$P(B = 1 A = 0) = 0.7$	$P(B = 1 A = 1) = 0.6$

Table 1: Example of conditional probability table of node B having node A is unique parent.

One of the most significant feature of Bayesian network is the notion of *conditional independence*. Simply speaking, for any node X in a Bayesian network, given the knowledge of node X 's parents, X is conditionally independent of all nodes that are not its children, or all its predecessors [9]. For example, in the Bayesian network in Figure 1, node C is conditionally independent of node A , when conditioned on node B being fixed. The possibility of exploiting conditional dependencies when computing the induced distribution of the Bayesian network is a powerful property since it simplifies the conditional probability table tremendously. For

example, the conditional probability table of node C , will not contain entries $P(C|A, B)$ since $P(C|A, B) = P(C|B)$, or node C is independent of A conditioned on B : $A \perp C|B$.

In the context of stress testing, Rebonato [18] suggests a subjective approach to constructing Bayesian networks. After carefully selecting a set of random variables as the nodes of the network, Rebonato proposes to subjectively connect the variables and assign the relevant conditional probability tables with the help of risk managers or other experts. Then with the inferred Bayesian network, reasoning about stressed events or simulation can be conducted. Please see [18] for details.

Our framework builds upon many of Rebonato’s intuitions but exploits our recent works on causality to address all the key problems, of which the subjective approach comes short. The subjective approach is handy under the condition of expert knowledge of the causal relationships of some variables. However, such reliance becomes unnatural when experts are confronted with random variables that are clearly beyond their expertise: for example, the relationship of unemployment and stock market performance, or more simply, the relationship of two random stocks. Therefore, instead of completely abandoning the role of data in the construction of Bayesian network, here we adopt learning algorithms that can learn both the structure and the conditional probability table of the Bayesian network from the data, which, in turn, can be further augmented by expert knowledge if deemed necessary.

2.2 Suppes-Bayes Causal Networks

Thus, unlike [18], our stress testing approach builds on the foundation of Suppes-Bayes Causal Networks (SBCNs), which are not only more strictly regularized than the general Bayesian networks but also enjoys many other attractive features such as interpretability and refutability. SBCNs exploit the notion of probabilistic causation, originally proposed by Patrick Suppes [23].

In [23], Suppes described the notion of *prima facie causation*. A *prima facie* causal relation between any event u and its effect v is verified when the following two conditions hold: (i) *temporal priority* (TP), i.e., any cause happens before its effect and (ii) *probability raising* (PR), i.e., the presence of the cause raises the probability of observing its effect.

DEFINITION 1 (PROBABILISTIC CAUSATION, [23]). *For any two events u and v , occurring respectively at times t_u and t_v , under the mild assumptions that $0 < P(u), P(v) < 1$, the event u is called a *prima facie* cause of v if it occurs before and raises the probability of v , i.e.,*

$$\begin{cases} (TP) & t_u < t_v \\ (PR) & P(v | u) > P(v | \bar{u}) \end{cases} \quad (1)$$

The notion of *prima facie* causality was fruitfully exploited for the task of modeling cancer evolution in [10, 14, 4], and the SBCNs were finally described for the first time in [3] and, later on, defined in [15, 16] as follow.

DEFINITION 2 (SUPPES-BAYES CAUSAL NETWORK). Let us consider an input cross-sectional dataset D of n Bernoulli variables and m samples, the Suppes-Bayes Causal Network $SBCN = (V, E)$ subsumed by D is a directed acyclic graph such that the following requirements hold:

[Suppes’ constraints] for each arc $(u \rightarrow v) \in E$ involving a *prima facie* relation between nodes $u, v \in V$, under the mild assumptions that $0 < P(u), P(v) < 1$:

$$P(u) > P(v) \quad \text{and} \quad P(v | u) > P(v | \neg u).$$

[Simplification] let E' be the set of arcs satisfying the Suppes’ constraints as before; among all the subsets of E' , the set of arcs E is the one whose corresponding graph maximizes the likelihood of the data and of a certain regularization function $R(f)$:

$$E = \underset{E \subseteq E', G=(V, E)}{\operatorname{argmax}} (LL(D|G) - R(f)).$$

Intuitively, the advantage of SBCNs over general Bayesian networks is the following. First, with Temporal Priority, SBCN accommodates the time flow among the nodes. There are obvious cases where some nodes occur before the other and it is generally natural to state that nodes that happen later cannot be causes (or parents) of nodes that happen earlier. Second, when learning general Bayesian networks, arcs $A \rightarrow B$ and $A \leftarrow B$ may sometimes be equally acceptable, resulting in an undirected arc $A - B$ (this situation is called Markov Equivalence [9]). For SBCNs, such a situation does not arise because of the temporal flow being irreversible [15, 16]. Third, because of the two constraints on the causal links, the SBCN graph is generally more sparse (has fewer edges) than the graph of general Bayesian networks with the final goal of disentangling spurious arcs, e.g., due to spurious correlations [13], from genuine causalities.

2.3 Machine Learning and Classification

Even if with SBCNs we typically obtain sparser DAGs than when we use Bayesian networks, the modelled relations both involve positive and negative financial scenarios, but only in the latter financial stress may arise. Thus, the extreme events which are of key relevance for stress testing are still rare in the data and unlikely to be simulated in naively generated stress scenarios by sampling from the SBCN directly. Therefore, in this work we improve this basic model with several key ideas of classic machine learning, namely, feature classification. Recall that, in stress testing, we wish to target the unlikely, but risky scenarios. Specifically, when generating random sample from a SBCN to obtain possible scenarios, each node in the SBCN can take any value in its support according to its conditional probability table, generating different branches of scenarios. To narrow down the search space, we can classify each possible branch as leading to *profitable* or *lossy* scenarios, and if, the branch is classified as profitable, then random sampling is guided to very likely avoid that branch, thus focusing on events and causal relations that can be adversarial and risky, though uncommon. In this way, computation can be reduced significantly to discover the extreme events (see the next Sections for details).

3. RESULTS AND DISCUSSION

In this Section we investigate the capabilities of our framework based on SBCNs to perform stress testing assessment. We first describe the adopted generative model and then present our algorithm and investigate its performance paying specific attention to the problem of false discoveries.

3.1 Simulating the Training Data

To assess the performance of the algorithm to learn the SBCNs and the quality of inferred Bayesian networks, a set of training data is developed with embedded causal relationships. If the algorithms, when performed on the training data are capable of accurately recover the causal relationships embedded in them, then such accuracy can be expected on real data.

To simulate the training data, we adopt a common stock factor model, the *Fama French Five Factor Model* [7], where the return of the asset is defined as follow:

$$r = R_f + \beta_1(K_m - R_f) + \beta_2SMB + \beta_3HML + \beta_4RMW + \beta_5CMA + \alpha. \quad (2)$$

In the equation, r is the return of the asset; R_f is the risk free return, usually measured in terms of government treasury returns; K_m stands for market factor, measured as value-weighted market portfolio, similar to stock indexes; SMB (Small Minus Big) stands for company size factor, measured by return on a diversified portfolio of small stocks minus the return on a diversified portfolio of big stocks; HML (High Minus Low) stands for company book-to-market (B/M) ratio factor, measured by difference between the returns on diversified portfolios of high and low B/M stocks, where B/M is the ratio between company's book value to market value; RMW (Robust Minus Weak) stands for company operating profitability factor, measured by the returns on diversified portfolios of stocks with robust and weak profitability and CMA (Conservative Minus Aggressive) stands for company investment factor, difference between the returns on diversified portfolios of low and high investment stocks, called conservative and aggressive [7].

To simulate the training data with embedded causal relationship, we linearly regress historical returns r , onto the five factors, and obtain the distribution of each factor coefficient and the empirical residual. We notice that a key characterization of a SBCN is an underlying temporal model of the causal relations depicted in the network, namely the temporal priority between any pair of nodes which are involved in a causal relationship. Therefore, the five factors described in our generative model are lagged with respect to the historical returns to comply with the temporal priority. Thus,

$$r_{i,t} = \sum_{i,j} \beta_{i,j} f_{j,t-lag} + \epsilon. \quad (3)$$

Then, the simple training data is simulated by randomly drawing the factor coefficients $\beta_{i,j}$ and residuals ϵ from the distribution we obtained from the linear regression, and apply these coefficients and residuals on a set of new factor data. Such historical data consists of daily series of five factors and returns of 10 portfolios also constructed by Fama French, and of 10000 days. We use the first 5000 for regression and the other 5000 for simulation.

In reality, many factors will present causal relationships among themselves. For example, some factor do not directly influence the asset, but affect the asset indirectly by affecting other factors. Therefore, the simulated training data can be complicated by embedding spurious relationships also among factors. We linearly regress some factors on the other factors and simulate the training data in the same way. The choice of factors is arbitrary. In this paper, as an example, we regress the other four factors SMB , HML , RMW and

CMA on the market factor K_m .

Therefore, the causal relationships which are described in the simulated training data can be simplified as shown in Figure 2.

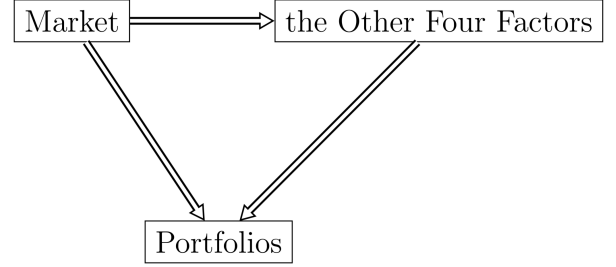


Figure 2: Example of causal relationships described in the simulated training data.

3.2 Learning the SBCNs

In this Section we show results on 100 independent random simulations generated on networks of 15 nodes, i.e., 10 stocks and 5 factors with the generative model discussed in the previous Section. Each node represents a Bernoulli random variable taking binary values in $(0, 1)$, where 1 represents the stock or factor going up and 0, the stock or factor going down. Specifically, the input of our learning task is a dataset D of $n \times m$ binary entries. Starting with such an input, we attempted to experiment with our learning algorithms previously described in [14] and [3]. In particular, as in [3], we lacked explicitly observed time in the data, which are only cross-sectional. To overcome this problem we imputed as a further input to our algorithm a topological ranking r providing information about the temporal priority among the nodes. In interpreting these experiments, we set ranking as a proxy of time precedence among the factors influencing the stocks, i.e., in our model factors can cause stock moves but not the other way around.

Algorithm 1 summarizes the learning approach adopted for the inference. Given the above mentioned inputs, Suppes' constraints are verified (Lines 3-8) to first construct a DAG. Then, the likelihood fit is performed by hill climbing (Lines 9-20), an iterative optimization technique that starts with an arbitrary solution to a problem (in our case an empty graph) and then attempts to find a better solution by incrementally visiting the neighbourhood of the current one. If the new candidate solution is better than the previous one, it is considered in place of it. The procedure is repeated until the stopping criterion is matched. In our implementation, the *!StoppingCriterion* occurs (Line 11) in two situations: (i) the procedure stops when we have performed a large enough number of iterations or, (ii) it stops when none of the solutions in $G_{neighbors}$ is better than the current G_{fit} , where $G_{neighbors}$ denotes all the solutions that are derivable from G_{fit} by removing or adding at most one edge.

For more information about the algorithm, also refer to [14, 3].

3.2.1 The Problem of False Discovery

We first test the performance of Algorithm 1 on a training data of 10 portfolios, 5 factors and 5000 observations. On

Algorithm 1 Learning the SBCN [3]

```
1: Inputs:  $D$  an input dataset of  $n$  Bernoulli variables and
    $m$  samples, and  $r$  a partial order of the variables
2: Output:  $SBCN(V, E)$  as in Definition 2
3: [Suppes' constraints]
4: for all pairs  $(v, u)$  among the  $n$  Bernoulli variables do
5:   if  $r(v) \leq r(u)$  and  $P(u | v) > P(u | \neg v)$  then
6:     add the arc  $(v, u)$  to  $SBCN$ .
7:   end if
8: end for
9: [Likelihood fit by hill climbing]
10: Consider  $G(V, E)_{fit} = \emptyset$ .
11: while !StoppingCriterion() do
12:   Let  $G(V, E)_{neighbors}$  be the neighbor solutions of
      $G(V, E)_{fit}$ .
13:   Remove from  $G(V, E)_{neighbors}$  any solution whose
     arcs are not included in  $SBCN$ .
14:   Consider a random solution  $G_{current}$  in
      $G(V, E)_{neighbors}$ .
15:   if  $score_{REG}(D, G_{current}) > score_{REG}(D, G_{fit})$ 
     then
16:      $G_{fit} = G_{current}$ .
17:   end if
18: end while
19:  $SBCN = G_{fit}$ .
20: return  $SBCN$ .
```

such settings, the algorithm was capable of recovering almost the whole set of embedded causal relationships with only 13 false negatives, roughly, 33% of total arcs; however, the number of false positives were unacceptably larger, reaching around 49% of the total causal arcs obtained, thus requiring more attention to how the model was regularized.

The explanation for this trend can be found in how the algorithm implements the regularization via Bayesian Information Criterion (BIC) [21], that is:

$$BIC = k \cdot \ln(N) - 2 \ln(L),$$

where k is the number of arcs in the SBCN (i.e. number of causal relationships), n is the number of observations of the data, and L is the likelihood. The algorithm searches for the Bayesian network that minimizes the BIC.

For large number of observations, the maximum likelihood optimization ensures that asymptotically all the embedded relationships are explored and the most likely solution is recovered. However, maximum likelihood is known to be susceptible to overfitting [9], especially when, as in our case, it deals with small sample size in the training data. Furthermore, in the training data, all the portfolios are assumed to depend on the same five factors, although with different coefficients, but very likely some portfolios will have very similar coefficients, resulting in co-movements across the portfolios. This co-movement will often induce correlations that affect the probability raising and thus the spurious prima facie causal relations, making these settings an interesting, and yet a very hard test case. See Figure 3.

3.2.2 Sample Size and Information Criterion

To reduce the spurious causalities, we recall some intrinsic properties of the information criteria. The Bayesian Information Criterion $BIC = k \cdot \ln(N) - 2 \ln(L)$, not only maximizes the likelihood, but also penalizes the complexity of

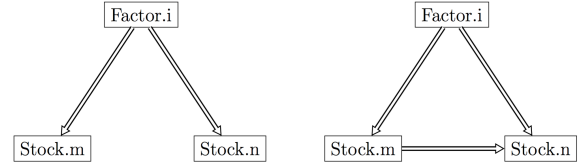


Figure 3: The figure simplifies the true causal relationships (on the left), and the spurious relationships (on the right) emerging from the simulated data.

the model by the term $k \cdot \ln(N)$. For small sample sizes, BIC is generally biased towards simple models because of the penalty. However, for large sample size, BIC is willing to accept complex models. For additional discussion, see the details in [9].

In our simulations we adopted a sample size of 5000 which is considerably large relative to the degree of freedom of the score function, therefore using BIC may lead to the inference of a relatively complex model with a number of unnecessary spurious arcs. Counter-intuitively, we could improve the solutions by using smaller sized data and letting the complexity penalty take a bigger effects in BIC score. This also addresses the nonstationarity in the data, an endemic problem for financial data. Following this intuition, we performed further experiments by reducing the original sample size of 5000 samples, which describes around 10 years of data, in turn to 250 and 500, and we observed a significant reduction in the number of false positives, to 38% and 40% of total arcs respectively. However, at the same time, because of smaller sample size, the number of false negatives inevitably increased to more than 50% of total arcs.

To reconcile this dilemma, we now next considered an alternative information criterion, the AIC, Akaike Information Criterion [1], defined as in the following:

$$AIC = 2k - \ln(L).$$

We notice that for AIC, the coefficient of k is set to 2, leading to definitely smaller factor than $\ln(N)$ of BIC when the sample size N is large. For this reason, AIC has the trend of accepting more complex models for given sample sizes than BIC. Applying AIC, the number of false negatives typically decreases, while the number of false positive gets larger.

3.2.3 Improving Model Selection by Bootstrapping

Up to this point, we described the different characteristics of two state-of-the-art likelihood scores with respect to the number of resulting false positive and false negative arcs. Specifically, we showed a trade-off where, because of their characteristics, the best results on large sample sizes is obtained using BIC, while for small sample sizes AIC is more effective, but neither of the two regularization schemes have a satisfactory trend. To improve their performance, now we propose to make use of a bootstrap [6] procedure for model selection.

The idea of bootstrap is the following: we first learn the structure and parameters of the SBCN as before, but we perform subsequently a resampling procedure where we sample with repetitions data from the dataset in order to generate a set of *bootstrapped datasets*, e.g., 100 times, and then we

calculate the relative confidence level of each arc in the originally inferred SBCN, by performing the inference from each of the bootstrapped dataset and counting how many times a given arc is retrieved. In this way, we obtained a confidence level for any arc in the SBCN.

We once again tested such an approach on our simulations and we observed empirically that the confidence level of spurious arcs are typically smaller than the confidence level for true causal relations. Therefore, a simple method of pruning the inferred SBCN to constrain for a given minimum confidence level is here applied. Such a threshold reflect the number of false positives that we are willing to include in the model, with higher thresholds ensuring sparser models. Here, we test our approach by requiring a minimum confidence level of 0.5, i.e., any valid arc must be retrieved at least half of the times.

We now conclude our analysis by showing in Tables 2 and 3 the contingency tables resulting from our experiments both for Algorithm 1 (TABLE 3) and for the standard likelihood fit method to infer Bayesian Networks (Table 2):

	BIC		BICBoot		AIC		AICBoot	
Sample	FP	FN	FP	FN	FP	FN	FP	FN
250	50.4	94.9	31.9	95.3	57.0	89.5	39.3	95.2
500	51.8	94.7	45.6	96.6	61.6	93.5	51.5	98.7
1000	55.2	93.5	47.6	94.5	63.7	85.5	50.4	95.2
2500	60.3	90.3	54.8	99.9	68.4	85.7	62.4	99.8
3500	62.5	84.2	58.0	94.3	69.5	85.7	63.7	96.8
5000	66.6	80.9	65.6	85.7	71.8	85.7	70.7	85.7

Table 2: Contingency Table of the Performances by standard Bayesian Networks of Different Information Criteria and Sample Sizes.

	BIC		BICBoot		AIC		AICBoot	
Sample	FP	FN	FP	FN	FP	FN	FP	FN
250	37.6	67.7	24.7	83.3	42.8	40.1	27.9	51.5
500	39.6	52.3	36.8	55.9	43.7	38.5	40.9	39.0
1000	43.1	39.1	40.7	47.8	47.9	37.6	43.9	38.0
2500	52.2	50.8	45.1	50.8	57.1	28.5	50.2	38.6
3500	48.7	33.3	48.7	33.3	58.8	28.5	57.1	28.5
5000	48.7	33.3	47.5	33.3	57.1	14.3	53.8	19.9

Table 3: Contingency Table of the Performances by Algorithm 1 of Different Information Criteria and Sample Sizes.

Table 3 presents the results in terms of false positives (FP) and false negatives (FN) by Algorithm 1 with the various methods on the training data with different information criteria, sample sizes, and whether Bootstrapping is applied. The trade-off between false positive rates and false negative rates usually is case-specific. We observe that, in general, the objective of such an approach is to correctly and precisely recover the true distribution underlying the training data. For this reason, unless differently specified for specific uses, there is not an overall preference toward either lower false positive or lower false negative. Therefore, we evaluate our methods by considering the sum of both false positive and false negative rates. This metric is biased toward a combination of relatively low FP and FN rather than the combination of very low FP and high FN and so on. By

analyzing the results shown in Table 3, we can clearly observe a trend where AIC with Bootstrapping on small sample datasets (i.e., 250) and BIC with Bootstrapping on large sample datasets (i.e., 5000) produces the best results, consistently with the discussion of the previous Section. Also, we observe that both for AIC without any bootstrapping on sample sizes of 250 and BIC without any bootstrapping on sample size of 5000, the false positive rates are reduced without a significant increase in the false negative rates.

We conclude by pointing out the significant increase in performance (both in term of FP and FN) when using SBCNs in place of standard BNs (compare Tables 2 and 3).

3.2.4 Assumption of Sparse Relationships

The resulting false positive rate may still seem relatively high. But, one important assumption is worth mentioning. In the training data, such high false positive rate derives from the fact that portfolios are dependent on the 5 common factors, which very likely will induce co-movements. However, in the real data, such nested dependencies do not always occur, while a feature of sparse relationships appears frequently, and portfolios depend on distinctively small sets of factors. This assumption of sparsity can significantly improve the performance of the algorithm.

To implement the assumption of sparsity, we deviate from the original Fama French five factor model. For simplicity, we generate data with sparse relationships using a random linear model with 10 factor variables and 20 stock variables. With 30% probability, each stock variable is dependent linearly on one of the 10 factor variables, so on average, each stock variable will be dependent on 3 factor variables, which will likely be distinct from the dependent factor variables of other stock variables. Then we sample factor variable data from normal distribution and compute the corresponding stock data using the linear model.

Implementing this sparsity on a new set of purely random training data we obtain with Algorithm 1 much better results, and, e.g., following the BIC with Bootstrapping method mentioned above, we obtain on small sample size (250 samples) 18.1% false positive and 39.2% false negative rates, while on large sample size (2500 samples), we obtain 50.2% false positive and 38.6% false negative rates.

Also the results for general BNs improve, but still being less effective than SBCNs.

3.2.5 Summary, ROC Space

We finally summarize in Figure 4 the results of this Section by interpolating and then smoothing out some kinks, and we obtain an ROC Space, whose x axis represent the False Positive Rate and y axis the True Positive Rate. ROC Space depicts the performance of the different methods we discussed on different sample sizes. By examining the plot, one can conclude that AICs generally have high true positive rates but also high false positive rates, as a result of its less stringent complexity penalty. In contrast, BICs generally have smaller false positive rates, but its true positive rates are also lower. Comparing the algorithms with and without bootstrapping, one can notice that the bootstrap procedure shifts the curves to the left. Still, the best performance lies in the data with the assumption of sparse relationships. Based on these results, we can conclude that with Bootstrapping and the assumption of sparse relationships, our algorithm is

capable of recovering accurately the causal relationships in the data.

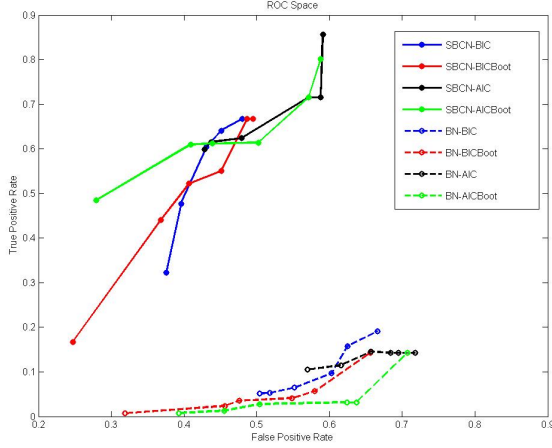


Figure 4: Performance in the ROC Space, depicts the trade-offs between false positive rates and true positive rates. The better results lie in the upper-left corner of the graph where false positive rate is low and true positive rate is high.

3.3 Performing Stress Testing

In this Section we present how to assess stress testing scenarios given the inferred Suppes-Bayes Causal Network and we present the results on the simulated data.

3.3.1 Risk Management by Simulations

After the inference of the SBCN, we perform Monte Carlo Simulation in the same way as conventional risk management, by drawing large number of samples to discover the worst 5% scenarios as the value at risk (VaR). Nevertheless, here in stress testing, we are targeting the most extreme events, which have very low but nonzero probability of occurrence. Thus, they still can occur, for example, the 2008 financial crisis or the most recent market reactions to *BREXIT*. Therefore, when drawing samples from the network, we would like to reject the normal scenarios, and place more importance on the extreme events. To achieve this goal, when conducting random sampling, we classify each possible branch as profitable or risky, and if the branch is classified as profitable, then we will avoid that branch.

Figure 5 represents a simple binary classification where for this factor only *Factor.i* with value 0 is considered risky and, hence, this scenario is the only one to be sampled. In this way, we target the extreme risky events and reduce computation. But, unlike conventional risk management, this approach does not allow us to estimate the probability of occurrence of the sampled extreme events, therefore we will not conclude a value at risk with certain confidence level.

The simple binary classification with certain features is a standard machine learning problem. Here we explore a simple solution of such a task based on decision trees [20]. A decision tree is a predictive models that maps features of an object to a target value of the object. Here, the features are the factors of interest, and the target value is whether

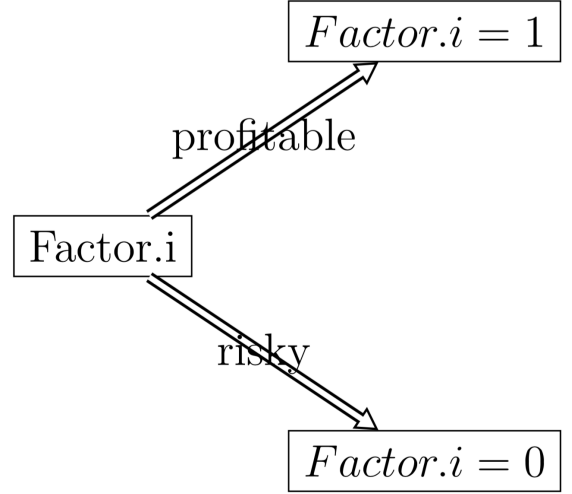


Figure 5: Risk classification in our SBCN.

the portfolio is prone to profit or loss. To perform classification, we first draw 1000 sample trajectories from the inferred SBCN. Then we construct a simple portfolio, which is long on all the stocks in the SBCN by the same amount, and calculate the Profit and Loss (P/L) of each observation. Here however, because the underlying SBCN depicts binary variables, exact Profit and Loss (P/L) statistics cannot be obtained. Instead, since the toy portfolio is long on all stocks by the same amount, the ratio of stocks that goes up is an approximate measure of risk. Of course, for continuous Bayesian network, Profit and Loss can be calculated directly. In the next step, we sort this measure, and denote the bottom-most 100 scenarios as risky, and the rest as profitable. The 100 ‘risky’ scenarios contain at least 7 stocks that fall. Then we consider 1000 samples each of them labeled as ‘risky’ or ‘profitable’. In our experiments, we used the R ‘tree’ package [19].

Using the SBCN learned from the simulated training data, we obtain the following decision tree shown in Figure 6.

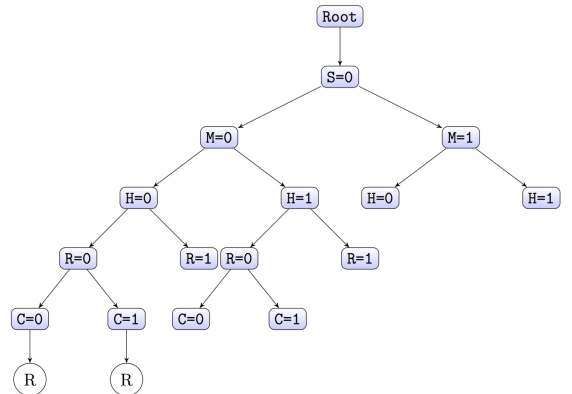


Figure 6: Decision tree obtained from the SBCN.

In the decision tree of Figure 6, *S* denotes factor *SMB*; *M* denotes Market *K_m*; *H* denotes *HML*; *R* denotes *RMW* and *C* denotes *CMA*. Here we show only the left part of

the entire computed decision tree, the subtree with $S = 1$ is omitted, since the entire subtree with $S = 1$ is classified as ‘Profitable’, which is not of interest for stress testing. In the tree, we identify two paths that are classified as ‘Risky’. Path $S = 0, M = 0, H = 0, R = 0, C = 0$ and Path $S = 0, M = 0, H = 0, R = 0, C = 1$. The paths classified are intuitive, since our example portfolio is long with equal amount invested over all 10 stocks. Since 10 stocks are generally positively dependent on the factors, most factors with 0 values will likely induce a ‘Risky’ path. For more complicated portfolios and real factors, such intuition cannot be easily found so we have to rely on the result of classification.

3.3.2 Scenario Generation and Results

Given the tree of Figure 6, we then used the `bnlearn` R package [22] to sample from the SBCN. Given the network, we can simulate random scenarios, however, we wish not to simulate all of them, which will prove to be inefficient, but following the informations provided by the classification tree we choose the configurations which are likely to indicate risk to drive our sampling. For instance, we may pick the first path in the tree, which is $S = 0, M = 0, H = 0, R = 0, C = 0$, and constrain the distribution induced by the SBCN. In order to avoid sampling the scenarios which are not in accordance with the path, we adjust the conditional probability table of the SBCN. Since we want paths with all five factors taking value 0, we set the conditional probability of these five factors taking value 1 to 0, and the conditional probability of factors taking value 0 to 1. In this way, the undesirable paths will be unlikely to be simulated, while the intrinsic distribution of how factors affect the stocks is still modeled. More sophisticated implementation based on this intuition are possible: e.g., using branch-and-bound, policy valuation, tree-search, etc, but we will leave this to further research.

Comparing the results of the simulations given the original SBCN and the one taking into account the decision tree, we show the distribution of the risk measure, the number of stocks that go up in the Figure 7.

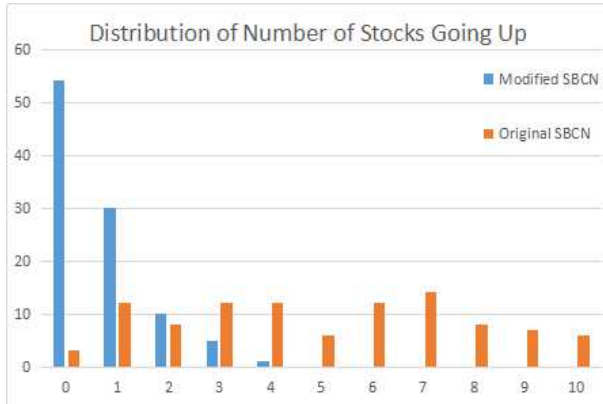


Figure 7: Distribution of Number of Stocks Going Up.

The number of stocks going up from 100 samples generated by the original SBCN is roughly evenly distributed. At the same time, the 100 samples generated by the modified SBCN contain no scenarios with more than 5 stocks going up, and 84 out of the 100 samples have at most 1 stock go-

ing up. We can clearly see that the modified SBCN places far more importance on the stressed scenarios, and in turn confirms the result of the classification algorithm by the decision tree. In this way, computational complexity involved in generating stressed scenarios can be improved tremendously. This kind of computational efficiency issues will be more critical when we move from a simple Bernoulli random variable to multi-categorical variables or continuous random variable. Therefore, with the same computing power, the modified SBCN makes it possible to generate more stressed scenarios, and observe how portfolios or other assets respond to stressed factors.

4. CONCLUSIONS

In summary, in this paper we develop a novel framework to perform stress testing combining Suppes-Bayes Causal Networks and machine learning classification. We learn SBCNs from data using Algorithm 1 and assess the quality of the learned model by switching information criteria based upon sample sizes and bootstrapping. We then simulate stress scenarios using SBCNs, but reduce computation by classifying each branch of nodes in the network as ‘profitable’ or ‘risky’ using classification trees. For simplicity, the paper implements SBCNs with Bernoulli variables and simulates data using Fama French’s Five Factor Model, but the logic of the problem is easily extended to deal with more practical situations. First of all, the SBCNs can accommodate more complicated variables (nodes). In addition to the factor based portfolios considered here, other factor models, or directly other financial and economic factors like foreign exchange rates, can also be included, and the accuracy of the model can ensure that the true causal relationships among the factors are discovered. In practice, variables like stock prices are continuous, thus, one can easily extend to these situations by adopting a hybrid SBCN, where the variables can take either discrete or continuous values, making it possible to represent precisely the values of the variables we are interested in.

To use the model, the role of experts is still important. After learning the SBCN from data and applying classification, we can identify a number of stressed scenarios. However, we expect that some of these to be unacceptable for various unforeseen reasons, e.g., as those known to domain experts. These scenarios may be thought of as highly stressed with respect to the corresponding portfolio but they could prove to be less useful in practice. Therefore, experts can select from the identified stressed scenarios only the plausible ones, and discard the ones deemed to be flawed. Even in this case, we can perform simulations following the selected stressed paths in the SBCN and observe the reactions of the portfolios in these stressed scenarios of interest, and thus adjust the portfolios based on the reactions. Another direct usage of our approach is when experts have a particular candidate stress scenario in mind, which can be justified a priori; in this case one can skip the process of classification and directly adjust the SBCN *mutatis mutandis*. Therefore, simulations of the adjusted SBCN will also offer the reactions of the portfolio to this particular stressed scenario.

We believe, based on our empirical analysis, that we have devised an efficient automated stress testing method using machine learning and causality analysis in order to solve a critical regulatory problem, as demonstrated by the algorithm’s ability to recover the causal relationships in the data,

as well as its efficiency, in terms of computation and data usage.

5. REFERENCES

- [1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.
- [2] N. Beerenwinkel, N. Eriksson, and B. Sturmfels. Conjunctive bayesian networks. *Bernoulli*, pages 893–909, 2007.
- [3] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti. Exposing the probabilistic causal structure of discrimination. *arXiv preprint arXiv:1510.00552*, 2015.
- [4] G. Caravagna, A. Graudenzi, D. Ramazzotti, R. Sanz-Pamplona, L. De Sano, G. Mauri, V. Moreno, M. Antonioti, and B. Mishra. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *PNAS*, 2016.
- [5] S. Claessens and M. A. Kose. Financial crises: Explanations, types and implications. *IMF Working Paper Series*, 2013.
- [6] B. Efron. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981.
- [7] E. F. Fama and K. R. French. Multifactor explanations of asset pricing anomalies. *The journal of finance*, 51(1):55–84, 1996.
- [8] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2010.
- [9] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [10] L. O. Loohuis, G. Caravagna, A. Graudenzi, D. Ramazzotti, G. Mauri, M. Antonioti, and B. Mishra. Inferring tree causal models of cancer progression with probability raising. *PloS one*, 9(10):e108358, 2014.
- [11] S. Manganell and R. F. Engle. Value at risk models in finance. *European Central Bank Working Paper Series*, 2001.
- [12] C. on the Global Financial System. Stress testing at major financial institutions: survey results and practice. 2005.
- [13] K. Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, 60(359-367):489–498, 1896.
- [14] D. Ramazzotti, G. Caravagna, L. O. Loohuis, A. Graudenzi, I. Korsunsky, G. Mauri, M. Antonioti, and B. Mishra. Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, 2015.
- [15] D. Ramazzotti, A. Graudenzi, G. Caravagna, and M. Antonioti. Modeling cumulative biological phenomena with suppes-bayes causal networks. *arXiv preprint arXiv:1602.07857*, 2016.
- [16] D. Ramazzotti, M. Nobile S, A. Graudenzi, and M. Antonioti. Learning the probabilistic structure of cumulative phenomena with suppes-bayes causal networks. *Submitted*, 2016.
- [17] S. Raychaudhuri, S. J. Mason, R. R. Hill, L. M. A. ũnch, O. Rose, T. Jefferson, and J. W. Fowler. Introduction to monte carlo simulation. In *2008 Winter Simulation Conference*, 2008.
- [18] R. Rebonato. *Coherent Stress Testing: a Bayesian approach to the analysis of financial stress*. John Wiley & Sons, 2010.
- [19] B. Ripley. *Tree: Classification and Regression Trees*, 2016. R package version 1.0-37.
- [20] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. 1990.
- [21] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.
- [22] M. Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.
- [23] P. Suppes. *A probabilistic theory of causality*. North-Holland Publishing Company Amsterdam, 1970.